PROJECT TITLE: **BOOK CLUSTERING.**

PROBLEM STATEMENT: Creating a clustering feature for readers by determining the most relatable book to a particular book.

PROBLEM STATEMENT:

Reading the wrong book would at times lose a reader’s interest. Hence, it is always important to choose the right book especially when the person is ignorant about reading. This project would generate few right choices of books for the user. The decision will be made on the input that the user provides.

* The search would be based on a specific author.

This can be used in a library where books can be easily organised based on similar context which will help the reader in finding any book.

FUTURE REVENUE:

The recommender can also be used at book exhibitions where the user can take suggestion on which book to buy.

The books must later be rated by the reader, which will not only give the most rated and least rated books but also will work as an initial feedback for the recommender.

ABOUT: The Gutenberg dataset consists of 3,036 books written by 142 authors. It is a subset of the Project Gutenberg corpus. Each file in the dataset corresponds to a particular book by a particular author.

STRUCTURE: It is a set of multiple .txt file. Each file has non-tabular data having size from 2kb- 1200kb.

**Total File size:** 1.12 GB

**Age of the data:** The dataset was formed on April 2014

**Seasonality:** Since we do not find the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly. Hence, we conclude that no seasonality as there is no time component in the dataset.

**Author:** Lahiri, Shibamouli

**Title:** Complexity of Word Collocation Networks: A Preliminary Structural Analysis

**Book title:** Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics

**Publisher:** Association for Computational Linguistics

**Data:** <https://drive.google.com/file/d/0B2Mzhc7popBga2RkcWZNcjlRTGM/edit>

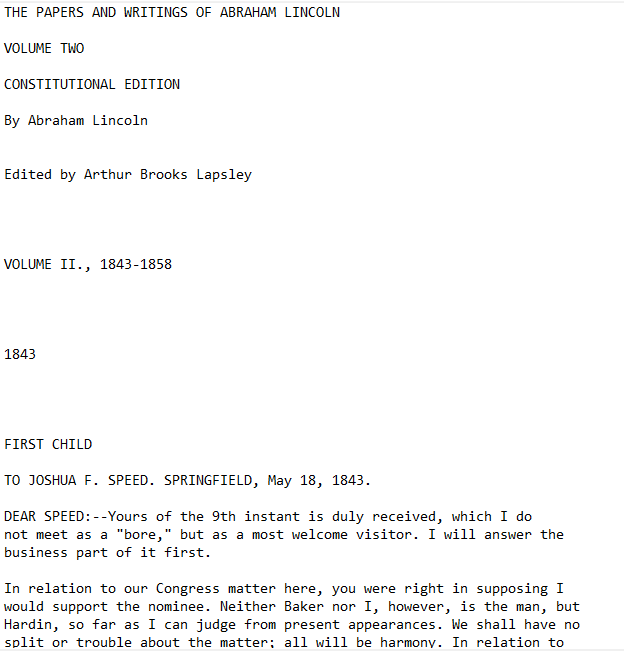
**Flow of the program 1:**

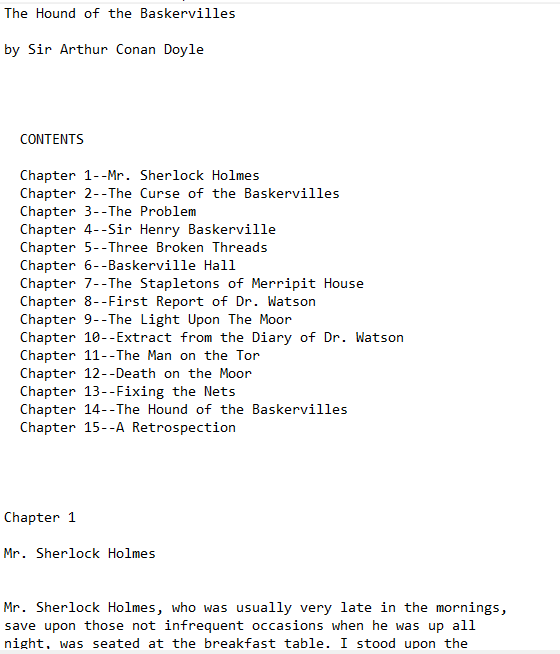
Reading the dataset -> Preprocessing the dataset -> Calculating TF-IDF values for each word in each document ->Converting to a dataframe structure -> Applying clustering Algorithm (K-means) -> Obtaining clusters to which each book should belong

**Flow of the program 2:**

Reading names of files -> Splitting it into author and book name -> Constructing a dataframe with key as author and value as list of book names -> Taking input from user -> Printing the list of books by the particular author.

SNIPPETS OF THE DATA:





Thank you!